Abstract

Large Language Models (LLMs) have achieved remarkable advances in reasoning, understanding, and generation across a wide range of natural language processing (NLP) tasks. However, their large scale imposes substantial computational and memory costs, posing challenges for efficient deployment, particularly in low-resource settings. Simultaneously, while retrieval-augmented methods have enhanced the factual grounding and interpretability of question answering (QA) systems, their effectiveness in low-resource languages remains limited by data scarcity and the absence of high-quality retrieval frameworks.

This thesis addresses these challenges through two complementary research directions. First, it introduces **LangCompress**, a language-aware model compression framework that integrates self-supervised instruction generation and vocabulary optimization to adapt LLMs for specific languages. LangCompress improves efficiency while preserving linguistic competence and can be seamlessly combined with existing pruning and quantization methods. Second, the thesis proposes a graph-based retrieval framework for multi-hop question answering that leverages Wikipedia's hyperlink structure to identify semantically connected evidence across documents. This framework enables efficient retrieval and reasoning in low-resource languages. To support this line of research, a generalizable dataset construction framework is developed, resulting in *VIMQA*, a Vietnamese multi-hop QA dataset designed to evaluate explainable and complex reasoning grounded in Wikipedia evidence.

Together, these contributions advance the development of retrieval- and compression-aware LLM systems that are efficient, interpretable, and inclusive across diverse linguistic environments.

Keywords: low-resource language, quantization, pruning, large language model, retrieval, question-answering